In the Claims:

| 1 | 1. A method for estimating a selectivity of a query containing at least one |
|----|---|
| 2 | column-associated condition related to column attributes of a relational database |
| 3 | table, the method comprising: |
| 4 | (a) generating a dataset by sampling a plurality of queries applied against the |
| 5 | database, wherein the dataset includes a plurality of query conditions and information |
| 6 | related to combinations of said query conditions, |
| 7 | (b) determining at least one regression function that reflects correlations |
| 8 | between particular query conditions based on said dataset, |
| 9 | (c) determining a table-specific estimate of a cardinality of a query based upon |
| 10 | the regression function serving as a data mining model. |
| 1 | 2. The method of claim 1, wherein step (c) further includes: |
| 2 | (c.1) selecting an access method for an incoming query from a plurality of |
| 3 | database access methods based upon the table-specific estimate for said incoming |
| | · |
| 4 | query. |
| 1 | 3. The method of claim 1, wherein said query includes column associated |
| 2 | conditions related to a plurality of tables, wherein step (c) further includes: |
| 3 | determining a table-combining cardinality estimate based upon said table- |
| 4 | specific estimate. |
| 1 | 4. The method of claim 1, wherein step (a) further includes: |
| | |
| 2 | (a.1) generating a dataset including queries q_j , $j=1,N$, wherein each query |
| 3 | includes a plurality of column-associated conditions c_{jk} , $k=1,M_j$, N , M being integer |
| 4 | variables, wherein step (a.1) further includes: |
| 5 | (a.1.1) storing a cardinality C of an elementary operation associated with a |
| 6 | column-associated condition c _{jk} , |
| 7 | (a.1.2) storing a count of query-qualifying database records reflecting the |

8

correlation between the database table column attributes referred to in each

- 9 elementary operation, 10 wherein step (c) further includes: 11 (c.1) calculating a cardinality estimate CE of said query with the following 12 formula: $CE = \sum_{i=1...I} f(Z_i)$ 13 wherein f (Z_i) is a regression function, CE is a total of correlations between the 14 15 plurality of combinations of elementary operations used in said sampled queries, and 16 Z_i is a frequency of occurrence for one or more column-associated conditions c_{ik} , and 17 wherein step (b) further includes: 18 (b.1) generating said regression function using said data mining model. 1 5. The method of claim 4, wherein step (c) further includes: 2 (c.2) estimating the cardinality of each of the plurality of column-associated 3 conditions c_{ik} referring to the same column using the data mining model. 6. 1 The method of claim 1, wherein step (c) further includes: 2 (c.1) training the model by using queries that include logical AND operators to 3 determine a correlation between corresponding column predicates. 7. 1 The method of claim 1, wherein step (c) further includes: 2 (c.1) transforming a query containing OR predicates to an equivalent query containing 3 AND predicates to simplify training of a model. 1 The method of claim 1, wherein step (c) further includes: 2 (c.1) normalizing the determined cardinality based upon a total number of
 - 9. The method of claim 1, wherein step (c) further includes:
- 2 (c.1) normalizing the cardinality associated with a sampled query with a size 3 of the database table when the query is sampled, and
- 4 (c.2) denormalizing a cardinality associated with a query for which a

rows in the database table.

3

5 cardinality is to be predicted with the size of the database table when the selectivity 6 for that query is predicted. 1 10. The method of claim 1, wherein step (b) further includes: 2 (b.1) using a subset of frequently used queries to determine said regression 3 function. 1 11. The method of claim 1, wherein step (b) further includes: 2 (b.1) repeatedly training said regression function with updated sampled data. 12. The method of claim 1, wherein step (a) further includes: 1 2 (a.1) sampling said queries via a tool based on a database optimizer. 13. The method of claim 1, wherein step (a) further includes: 1 (a.1) determining cardinalities for individual table columns via a database 2 3 statistics tool, and 4 (a.2) mapping queries that include a plurality of logical AND operators to 5 corresponding cardinality based regression formulae. 1 14. The method of claim 1, wherein step (a) further includes: 2 (a.1) mapping queries that include at least one of an inner join and an outer join to corresponding regression formulae based on at least one of cardinality and 3 4 selectively operations. 1 15. A method for determining an access plan for a database query 2 compatible with data mining based database access control comprising: 3 (a) selecting a regression function for use with said query, 4 (b) determining a number of qualifying records for said query via said 5 regression function, and 6 (c) selecting an access method for accessing the database from a plurality of

different access methods based upon the determined number of qualifying records.

| 1 | 16. A computer system for estimating a selectivity of a query containing at |
|----|--|
| 2 | least one column-associated condition related to column attributes of a relational |
| 3 | database table, the system comprising: |
| 4 | a sampling module for generating a dataset by sampling queries applied |
| 5 | against the database, wherein the dataset includes a plurality of query conditions and |
| 6 | information related to combinations of said query conditions, |
| 7 | a regression module for determining at least one regression function that |
| 8 | reflects correlations between particular query conditions based on said dataset, |
| 9 | a processing module for determining a table-specific estimate of a cardinality |
| 10 | of a query based upon the regression function serving as a data mining model. |
| | |
| | |
| 1 | 17. The system of claim 16, wherein the processing module selects an |
| 2 | access method for an incoming query from a plurality of database access methods |
| 3 | based upon the table-specific estimate for said incoming query. |
| 1 | 18. The system of claim 16, wherein said query includes column |
| 2 | associated conditions related to a plurality of tables, and wherein the processing |
| 3 | module determines a table-combining cardinality estimate based upon said table- |
| 4 | specific estimate. |
| 7 | specific estimate. |
| 1 | 19. The system of claim 16, wherein the sampling module further |
| 2 | comprises: |
| 3 | a dataset module for generating a dataset including queries q_j , $j=1,N$, |
| 4 | wherein each query includes a plurality of column-associated conditions c_{jk} , $k=1,M_j$ |
| 5 | N, M being integer variables, wherein said dataset module further comprises: |
| 6 | a first storage module for storing a cardinality C of an elementary operation |
| 7 | associated with a column-associated condition c_{jk} , |
| 8 | a second storage module for storing a count of query-qualifying database |
| 9 | records reflecting the correlation between the database table column attributes |

10 referred to in each elementary operation,

wherein the processing module further comprises:

an estimation module for determining a cardinality estimate CE of said query

13 with the following formula:

14
$$CE = \sum_{i=1,..L} f(Z_i)$$

- wherein $f(Z_i)$ is a regression function, CE is a total of correlations between the
- 16 plurality of combinations of elementary operations used in said sampled queries, and
- Z_i is a frequency of occurrence for one or more column-associated conditions c_{jk} , and
- wherein the regression module further comprises:
- a function module for generating said regression function using said data
- 20 mining model.

- 1 20. The system of claim 19, wherein the processing module estimates the
- 2 cardinality of each of the plurality of column-associated conditions cik referring to the
- 3 same column using the data mining model.
- 1 21. The system of claim 16, wherein the processing module trains the
- 2 model by using queries that include logical AND operators to determine a correlation
- 3 between corresponding column predicates.
- 1 22. The system of claim 16, wherein the processing module transforms a
- 2 query containing OR predicates to an equivalent query containing AND predicates to
- 3 simplify training of a model.

| l | 23. The system of claim 16, wherein the processing module normalizes the |
|---|--|
| 2 | determined cardinality based upon a current total number of rows in the database |
| 3 | table. |
| | 24 The section of the in-16 subsects the section we delegate the section of the s |
| l | 24. The system of claim 16, wherein the processing module normalizes the |
| 2 | cardinality associated with a sampled query with a size of the database table when the |
| 3 | query is sampled, and denormalizes a cardinality associated with a query for which a |
| 1 | cardinality is to be predicted with the size of the database table when the selectivity |
| 5 | for that query is predicted. |
| 1 | 25. A program product apparatus having a computer readable medium with |
| 2 | computer program logic recorded thereon for estimating a selectivity of a query |
| 3 | containing at least one column-associated condition related to column attributes of a |
| 1 | relational database table, said program product apparatus comprising: |
| 5 | a sampling module for generating a dataset by sampling queries applied |
| 5 | against the database, wherein the dataset includes a plurality of query conditions and |
| 7 | information related to combinations of said query conditions, |
| 3 | a regression module for determining at least one regression function that |
|) | reflects correlations between particular query conditions based on said dataset, |
|) | a processing module for determining a table-specific estimate of a cardinality |
| l | of a query based upon the regression function serving as a data mining model. |
| l | 26. The program product of claim 25, wherein the processing module |
| 2 | selects an access method for an incoming query from a plurality of database access |
| 3 | methods based upon the table-specific estimate for said incoming query. |
| l | 27. The program product of claim 25, wherein said query includes column |
| 2 | associated conditions related to a plurality of tables, and wherein the processing |
| 3 | module determines a table-combining cardinality estimate based upon said table- |
| 1 | specific estimate. |
| l | 28. The program product of claim 25, wherein the sampling module |

further comprises:

a dataset module for generating a dataset including queries q_j , j=1,...N,

4 wherein each query includes a plurality of column-associated conditions c_{jk} , $k=1,...M_j$,

5 N, M being integer variables, wherein said dataset module further comprises:

a first storage module for storing a cardinality C of an elementary operation

7 associated with a column-associated condition c_{jk} ,

8 a second storage module for storing a count of query-qualifying database

9 records reflecting the correlation between the database table column attributes

10 referred to in each elementary operation,

wherein the processing module further comprises:

an estimation module for determining a cardinality estimate CE of said query

with the following formula:

$$CE = \sum_{i=1...L} f(Z_i)$$

11

wherein $f(Z_i)$ is a regression function, CE is a total of correlations between the

plurality of combinations of elementary operations used in said sampled queries, and

17 Z_i is a frequency of occurrence for one or more column-associated conditions c_{jk} , and

wherein the regression module further comprises:

a function module for generating said regression function using said data

20 mining model.

- 1 29. The program product of claim 28, wherein the processing module
- 2 estimates the cardinality of each of the plurality of column-associated conditions cik
- 3 referring to the same column using the data mining model.
- 1 30. The program product of claim 25, wherein the processing module
- 2 trains the model by using queries that include logical AND operators to determine a
- 3 correlation between corresponding column predicates.